TEC-0092

Task-Driven Active Vision for Security and Surveillance

H. Keith Nishihara
Richard Marks
Stanley J. Rosenschein

J. Brian Burns
Philip Kahn

Teleos Research
2465 Latham Street, Suite 101
Mountain View, CA 94040

August 1998

Prepared for:
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714

Monitored by:
U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road
Alexandria, VA 22315-3864

US Army Corps of Engineers
Topographic Engineering Center

T

E

C

U.S. ARMY TOPOGRAPHIC ENGINEERING CENTER
TEC

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1998	Technical October 1994-September 1995

4. TITLE AND SUBTITLE

Task-Driven Active Vision for Security and Surveillance

5. FUNDING NUMBERS

DACA76-93-C-0017

6. AUTHOR(S)

H. Keith Nishihara J. Brian Burns Philip Kahn
Stanley J. Rosenschein Richard Marks

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Teleos Research
2465 Latham Street, Suite 101
Mountain View, CA 94040

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Advanced Research Projects Agency
3701 North Fairfax Drive, Arlington, VA 22203-1714

U.S. Army Topographic Engineering Center
7701 Telegraph Road, Alexandria, VA 22315-3864

19. SPONSORING / MONITORING AGENCY REPORT NUMBER

TEC-0092

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This annual report details the progress made in the development of computer vision and visual attention mechanisms for the support of Security and Surveillance applications. Progress is reported in two areas: real-time methods for discriminating moving shapes against moving backgrounds; and object recognition using consensus-based techniques to increase robustness and computational efficiency.

14. SUBJECT TERMS

Security and Surveillance, Active Vision,
Figure-Ground Discrimination, Object Recognition

15. NUMBER OF PAGES

31

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Contents

List of Figures

PREFACE

This research is sponsored by the Defense Advanced Research Projects Agency (DARPA) and monitored by the U.S. Army Topographic Engineering Center (TEC), under Contract DACA 76-93-C-0017, titled, "Task-Driven Active Vision for Security and Surveillance." The DARPA point-of-contact is Dr. Pradeep Khosla and the TEC Contracting Officer's Representatives are Ms. Lauretta Williams and Mr. Thomas Hay.

1 Introduction

This is the second annual report on research accomplished by Teleos Research on a three-year contract supported by DARPA's Real-Time Planning and Control Program.

1.1 Motivation

Computer perception applied in the security and surveillance domain has a wide range of immediate governmental and commercial applications that include: law-enforcement and security (e.g., detection of public criminal activities, such as drug dealing on street corners, building surveillance, detection of loitering and parking lot security), nuclear storage warehouse security, and consumer mobility pattern analysis in a store (e.g., to detect shoplifters, optimize product placement for consumer traffic patterns). Current automated security systems are prone to high false alarm rates and often the only acceptable solutions require direct monitoring by human personnel.

A key capability required by the above applications is the ability to automatically, rapidly and consistently recognize objects and temporal events observed under natural viewing conditions. This requirement is made difficult by the real-time nature of the task and the complexity of finding and analyzing object images when they are undergoing articulated motions under varying lighting and against complex, possibly moving, backgrounds. So far, current vision-based technologies have been insufficient to meet these demands.

To address these considerations, Teleos has directed its perception research effort towards time-critical measurement, figure detection, and control tasks. From the research perspective the Security and Surveillance (S&S) domain is a visually rich, but otherwise restricted, application domain that is vital for guiding and evaluating a core research effort of this type. S&S is a good environment in which to test task-directed vision techniques. It supports perceptual tasks over a broad range of difficulty. These include detection of new or reappearing objects, classification of movement patterns, detection of common destinations, detection and tracking of motion in visual or IR imagery with an active head, and discrimination of humans using size, shape, color, texture, and motion cues. S&S problems often require visual strategies in order to perform well, and goal-

directed attentional mechanisms are key in all but simple cases.

1.2 Overall Research Plan

The work under this contract is organized into three components: (1) The first emphasizes real-time perception mechanisms involving motion-based detection of figures. It involves tracking those figures, once detected, with a camera having pan-tilt-zoom control. The bulk of this research was completed during the first contract year. Additional enhancements to the figure discrimination and tracking algorithms were added during the current program year. (2) The second component of the research effort focuses on reacquisition mechanisms that allow the tracking of multiple targets undergoing various kinds of occlusion and for discriminating people from other moving agents. The bulk of this work was completed during the current program year and a detailed presentation on the results is given in Section 3. (3) The third component will focus on recording and classifying tracked agents based on their motion characteristics. The results of this work will be reported in the final report for this contract. Figure 1 illustrates the composition of these modular components.

1.3 Document Organization

This Annual Report details progress that Teleos has made in the development of computer vision and visual attention mechanisms for the support of an S&S-directed vision and planning system. Section 2 presents a summary of research and technology transfer activities accomplished during the program year. Section 3 gives a more detailed presentation of the theoretical results obtained for figure recognition.

2 Summary of Work Done in this Period

The major visual perception capabilities relevant to security and surveillance addressed in this research program are the detection of human subjects, the tracking of their motion

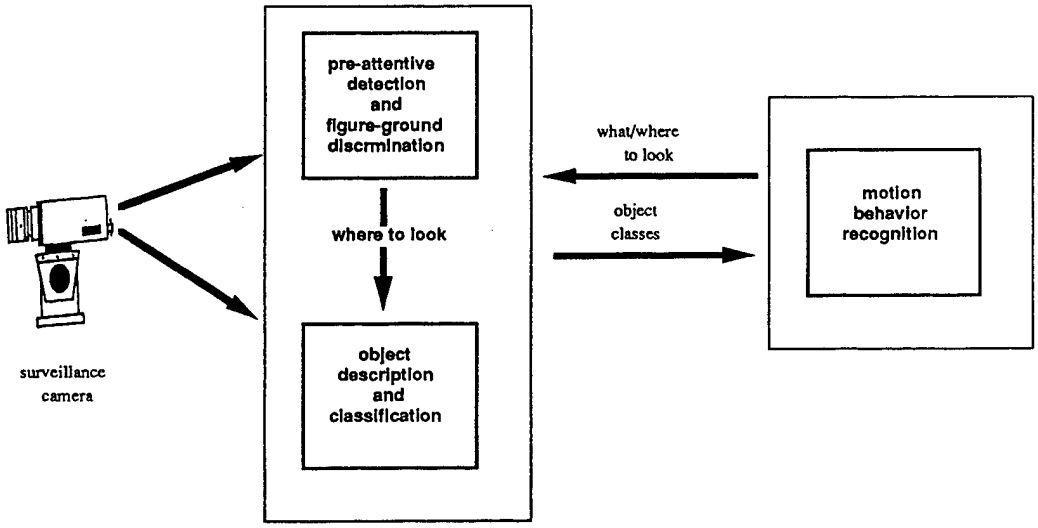

Figure 1: Component Technologies. Event perception is a process that naturally modularizes into three primary subtasks—detection and tracking; object classification; and temporal event recognition. Research sponsored under this DARPA Real-Time Planning and Control contract develops theory and technology in each of these areas. These component technologies are used to explore and evaluate an integrated system for use in security and surveillance applications.

and the re-detection of specific, previously monitored, subjects. Research begun in the previous annual reporting period was continued with work focused in two major areas: The development of robust techniques for discriminating a figure from its background in the presence of camera motion and unconstrained figure movements; and the development and testing of a consensus-based recognition theory suitable for classification and reacquisition of tracked figures.

Highlights of the work performed in each of these areas is summarized in this section.

2.1 Motion-based Figure-ground Discrimination

Automatic detection of object candidates (*figures*) from the background image prior to recognition is an important first step towards accomplishing recognition under dynamic time-critical conditions.

This figure extraction process helps to:

1. Allocate processing resources efficiently through attentional control
2. Normalize features by providing position and size estimates
3. Reduce interference in statistics by restricting where objects can be
4. Enable operation in dynamic domains where tracking is essential

By isolating a figure significantly smaller than the whole image, the resources of the recognition process can be employed more efficiently. More processing can be performed in the area likely to contain an object of interest.

The isolated figure has a size, position and perhaps a discernible orientation (such as an upright figure of a person). Increased efficiency can be achieved by using knowledge of the figure's geometry to normalize geometric features with respect to a certain change in view. Isolating the figure also can reduce interference from background data when computing color and texture statistics that are useful as recognition features.

Finally, figure extraction may be particularly feasible in many of the dynamic, time-critical domains that require it. For example, when the time constraints are caused by the fact that the object of interest is in motion, the differential image motion exhibited by the object's projection gives strong evidence of its position and extent. Figure 2 shows an example of a complex articulating figure (a person) extracted from the background via motion analysis. These frames are from a large sequence in which the Teleos-developed extraction process performs well.

2.1.1 Positive and Negative Criteria

The task of detecting figures moving against a background can be accomplished using positive information from the figure such as by tracking surface patches or other features on the object. Another approach is to model the more uniform motion of the background and detect areas in the image that do not move with the background. This latter approach can be thought of as using negative criteria to detect figures.

There are several advantages to using negative criteria. First, background motion is much easier to model than are the motions associated with a complex articulating figure. Second, objects moving faster than the vision system's ability to measure can still be detected as long as the background velocity can be handled. Third, negative criteria algorithms generally require less computation and the underlying theories are simpler.

The primary weakness associated with negative criteria is its inability to discriminate one moving figure from another. This leads to confusion anytime the tracked figure gets near another moving object.

Positive criteria, on the other hand, makes use of various similarity measures, such as color, texture or feature geometry, to maintain track on the same object over time.

A system integrating higher performance negative criteria with more selective positive criteria will ultimately provide the most robust operation. This research program develops an attentional and tracking mechanism based on the use of negative criteria. It then develops positive criteria methods for accomplishing reacquisition and object recognition.

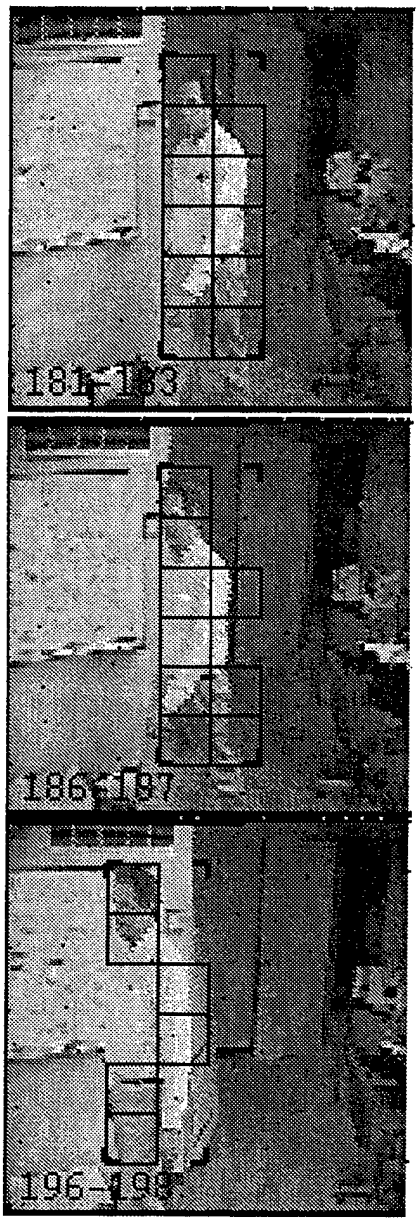

Figure 2: Figure Detection. Frames from a video sequence showing movement-based detection of an object (black squares). This *negative criteria* figure-ground discrimination algorithm allows detection of moving bodies against a moving background as occurs when the camera is in motion.

2.1.2 Tracking Using Negative Criteria

During the past program year, Teleos carried out research on negative criteria approaches for discriminating moving figures against a moving background as would be the case when the surveillance camera is in motion seeking targets or while tracking them. The real-time computation of image flow fields was assumed to be available as is the case with Teleos' AVP-100 stereo and motion measurement platform.

The negative criteria algorithm uses a single camera to measure optical motion densely over the camera's image. The velocity of the background motion is estimated by finding a peak in a histogram of the velocities in the dense motion field. This background velocity is nulled out in a correlation step that compares successive images. Figures with differential motion relative to the background pop out as poorly correlated regions in this comparison step. The locations of these regions are measured and reported as detected figures.

This basic design has been further enhanced to allow the detection of very slowly moving objects by freezing the reference image and updating it only when motion between the reference and current live image become too large to handle. A further enhancement maintains a representation of figure position when no figure motion is occurring.

This algorithm has been tested on video sequences of subjects walking in an indoor environment. It also has been interfaced to a commercial pan-tilt-zoom security camera and used to demonstrate active camera control in tracking people in indoor and outdoor settings.

2.2 Consensus-based Recognition

Two important recognition tasks in security and surveillance are (1) the re-acquisition of a tracked figure after some interruption, such as visual occlusion, and (2) the identification of the tracked figure given previously stored visual information. In both cases, the recognition module should have the following properties:

1. *Image-based*: The system should be able to recognize an object in a novel image sequence given images of the object previously taken from a sampling of views

2. *Dependable*: Given only a few frames of the novel image sequence, the system should be able to confidently determine whether or not the familiar object is present

3. *Stable performance*: Recognition should be possible under a range of viewing conditions, such as varying view, object scale and lighting

4. *Real-time:* Object identification should be completed within the time interval that the subject is being tracked. In the context of tracking human subjects, this should occur within a few seconds

Our previous efforts towards recognition focused on the extraction of human figures from backgrounds and the detection of salient figure parts, such as limbs for full-figures and facial features for faces. Principled figure extraction is an important first step towards confident recognition. In addition, the extraction and matching of salient parts is a useful approach to the detection of full, human figures undergoing large limb articulations.

The research of this past year complements the previous research by focusing on the detection of human faces. Our general method also can be used for the recognition of any other semi-rigid object. It meets the four requirements listed above. The algorithm involves three steps: (1) stable, efficient encoding of local image patches, (2) local patch matching via rapid indexing, and (3) global consensus of object match though a robust voting process. The image encoding method uses local moments of brightness and texture orientation, that ensures that the system is both image-based and stable with respect to large changes in viewing conditions. The stability of the local representation, together with the robust global consensus method, makes the overall algorithm potentially very dependable. Finally, each of the three steps can be implemented very efficiently, giving the system real-time capability.

A specific design based on this approach and experiments demonstrating the design are described in detail in Section 3.

2.3 Technology Transfer

Results of this research effort were shared with members of the Real-Time Planning and Control Program (RTPC) at PI meetings held 5-6 October, 1994 in the Washington, D.C. area, and more recently on 7-8 September, 1995 in Atlanta, GA.

Teleos Research also prepared and operated a live demonstration of RTPC funded research results at DARPA's Software and Intelligent Systems Technology Office Symposium held in the Washington D.C. area 26-31 September, 1995. This demonstration highlighted collaborative work between SRI and Teleos and brought together results from several DARPA supported programs including RTPC, Image Understanding, and Unmanned Ground Vehicle.

RTPC-sponsored research results have been run as an application layer on top of the AVP-100 system, a low-cost, commercially available, real-time visual measurement platform. This system including figure tracking algorithms, was delivered to four research laboratories during this annual reporting period. More widespread distribution of this technology is anticipated for the coming year.

Teleos also is actively working to transfer its RTPC-sponsored technology into practical commercial applications. Collaborations with several major companies in the security and video teleconferencing areas are underway.

3 Consensus-based Recognition

The study descibed in this section is an exploration of the fruitful combination of two useful methods in recognition: consensus or voting-based approaches and moment-based representations. The basic idea is first demonstrated using moments of the image brightness on the detection of 3-D objects undergoing 6-D variations in position and orientation (*pose*). The idea is then extended to handle large variations in light source using moments of local texture orientation. This idea also is demonstrated on real image data.

3.1 Consensus Methods

In consensus-based recognition, correspondences between localized parts of an object model and localized parts of the input image are formed, and each such local match votes for the object and object poses that are consistent with it. Detection occurs if there is a large enough accumulation of votes for some object and pose. This is the basic approach associated with generalized Hough transforms [1] and geometric hashing [2], both of which have been used to detect 3-D objects. The advantages of these techniques are simplicity and robustness with respect to large corruptions or loss of data such as by occlusion. Current 3-D recognition designs based on this approach are limited either by dependence on special, potentially difficult-to-detect local image features, such as line junctions [1], [2] and elliptical arcs, or image features that are too indistinct and ubiquitous, such as edge points. In the latter case, the system can be plagued by too many local feature correspondences to process (millions or billions), and too much clutter in the space or hash table in which the vote cluster detection is performed. Methods of detection by voting become much more efficient if the local features are more uniquely characterized by higher dimensional descriptions [3], and the process is more broadly applicable if the local feature representation used is more general than the detection of specific structures (such as line junctions and ellipses).

3.2 Moment Representation

Ideally, local feature representation would be a simple function of the original image data and the whole object would be modeled by processing images from distinct views, as in Figure 3. In the moment-based approach to recognition, the image is filtered by a set of 2-D functions that represent or are related to the moments of the brightness distribution in the image. These techniques include traditional moment methods [4] and [5], as well as current methods using steerable filters ([6], [7], [8]), and are related to techniques using Gabor wavelets [9]. The steerable filters currently used are the derivatives of the Gaussian taken at various scales [10], which are actually linear combinations of the brightness moments weighted by Gaussians at given scales [11]. Figure 3 shows all of the first and second derivatives of the 2-D Gaussian at two different scales, separated by a half-octave. Moments provide a relatively informative representation of the image that can be normalized with respect to changes in contrast and some image distortions. They

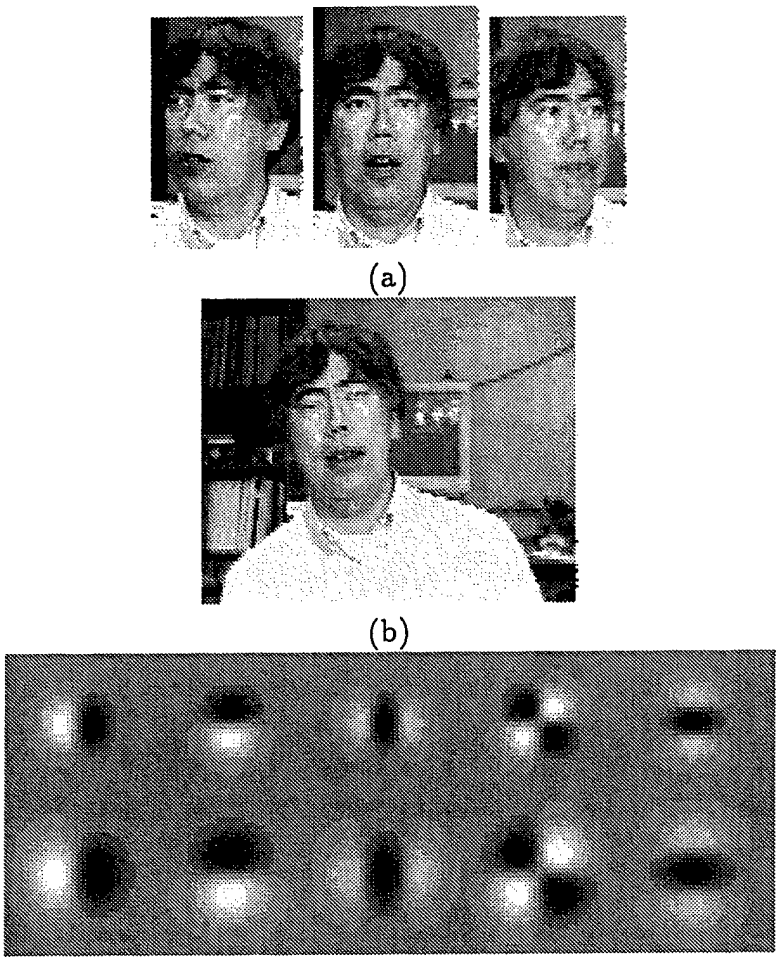

Figure 3: Moment Representation. Moments based on derivatives of Gaussian smoothed image patches provide a relatively informative representation of the image that can be normalized to minimize effects of changes in contrast and some image distortions. Tests of a view-based approach using a moment representation are illustrated here. (a) Image samples representing the subject from distinct views. (b) A sample from a 150 frame sequence of the subject talking, blinking and rotating, used to test the recognition system. (c) Derivatives of the 2-D Gaussian at different scales: the bottom row is one half-octave larger in scale than the top, and the derivatives are (left-to-right) g_x, g_y, g_{xx}, g_{xy} and g_{yy}.

also are a simple, general basis for representation, requiring only sample images of the object to generate a model (Figure 3). In addition, if a relatively small series of moments are used for each local patch, the image processing and matching of the moment features can be made fast on standard processors.

One shortcoming of moment matching is the stability of the moments with respect to occlusions, image clutter and other disruptions. Because of this, moments have been traditionally used to detect objects that are isolated and often in silhouette form [4], [5]. Recently, Rao and Ballard [7] have used methods of robust feature vector matching to improve this, going far to demonstrate the discrimination potential of moments from a single patch. However, their design is still very sensitive to certain disruptions; it accepts too many false positives, and it may have high storage costs. To get the large number of Gaussian derivatives (45) required for their robust method, the image patch encoding uses nine derivatives measured over five octaves of scale. If the encoding is centered and scaled so that the image support for the largest scale is largely within the boundary of the object (i.e., extraneous data have low impact), then the support region for the smallest scale occupies only 1/256th of the object's region, and three-fifths of the features have support regions that are only one-sixteenth of the object region. This should make the system very sensitive to occlusion or other changes in a small, central area of the image patch.

Also, with only forty-five measurements and the very tolerant match acceptance required for robustness, the moments of a single image patch alone have a false positive rate that can be improved on: multiple false points in a single image are said to have at least 0.9 correlation with a given reference patch, and a seventy percent discrimination rate across a set of rigid objects is reported. Finally, Rao and Ballard's robust feature methods also may incur an efficiency penalty: to handle random perturbations of a forty-five component moment vector, the feature indexing system may have to store a lot of the vector variations.

Instead of depending on measurements from a single image patch to represent and detect an object, it seems better, in terms of occlusion insensitivity and processing complexity, to use localized measurements from multiple patches distributed about the object's image and of different sizes. The multiple matches of these individual patches can be used to detect and localize an object via the consensus or voting methods discussed above. This potentially fruitful combination of methods is stressed and demonstrated here: the approaches of moment representation and vote clustering are complementary.

The use of multiple, local moment-based matches also is a basis for the method of [8]; however, they only use a few, manually selected local patches (five) and no uniform method of determining large clusters of consistent matches. Rao and Ballard [7] also discuss using multiple patches, but again, only in a sparse sampling and without a method of detection by integrating multiple match results under arbitrary variations in 3-D position and orientation. A larger number of locally encoded patches is considered in [9], but this work emphasizes an iterative match refinement technique. Their method is potentially useful for verifying and improving the pose estimate after the object has been roughly detected and located.

Our basic idea of local moment matching and global voting-based cluster detection has been applied to moments of brightness and local texture orientation. The latter shows promise in the presence of large, complex changes in lighting.

3.3 Brightness Moments

3.3.1 Basic Design

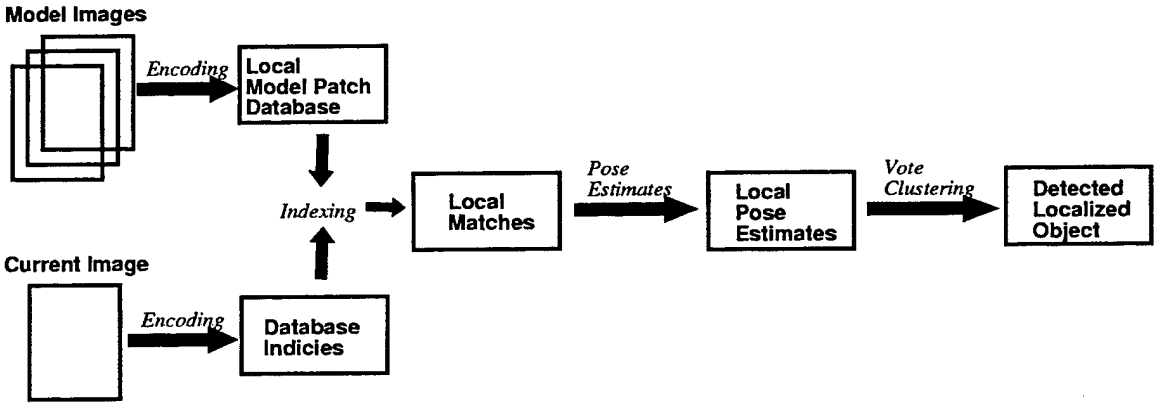

Figure 4: Voting Architecture Design. Basic design of recognition system that combines voting for object detection and moments for local feature encoding and indexing.

Our approach has the following steps (Figure 4.) Prior to recognition, images of the object are taken from a sampling of 3-D views. For each of these model images, local

image patches are automatically selected at various image positions and scales. Each patch is encoded by a set of Gaussian derivatives. These features are normalized with respect to certain image changes and transformed in a way to be optimum for patch discrimination. Information about each patch is stored in a model data base indexed by the features.

During recognition, patches in the input image are selected and encoded in a similar fashion. Each encoded patch is used as an index into the data base to retrieve potentially matching model patches, that are then tested for quality of match. Each matching pair of input and model patches then votes for the approximate object pose consistent with their match, incrementing an accumulator cell associated with the pose. After voting, the cell with the largest vote count is selected, and, if the count is above some threshold, the associated object pose is returned.

3.3.2 Local Representation

A local image patch is represented by a set of derivatives of Gaussians. The measurements range in Gaussian scale and order of derivative. As discussed above, it may not be advantageous to use too large a range in scales. In addition, since our system localizes the object by consensus of multiple matches, it is not essential that any given local patch be uniquely matched—as long as the overall processing time is reduced enough by initial patch matching. Thus, only a small number of scales and derivatives may be required. For the experiments presented here, two scales separated by an octave and all the first and second derivatives in the x and y directions were used; this creates ten measurements per patch.

The measurements are normalized with respect to changes in contrast and rotation in the image using the gradient of the larger scale: the responses are divided by the gradient magnitude, and, as in [6], rotated so that the larger scale gradient is parallel to the x-axis. This gives us eight remaining measurements to use as features. Since all the odd derivatives and all the even derivatives are dependent, the feature space is transformed using principal component analysis. For patch samples from the model images used in the experiments, the resulting eight features have a covariance matrix that is approximately the identity matrix.

Experiments here and elsewhere [6], [7] show that Gaussian derivatives have stability with respect to view changes, and are stable enough to be useful for discrimination purposes. In [6], good matches are demonstrated after 3-D rotations of up to 22.5 degrees. Experiments here have shown stability with respect to scale changes. After changing the image scale by as much as 20 percent, the resulting feature vector's average distance from the unscaled patch's vector is very small relative to the distances between randomly paired patches. In one experiment, a threshold was selected such that 50 percent of the distances between scaled patches and their unscaled counterparts were less than this threshold. (A sample of 20,000 patches were used.) Of a population of 9,000 randomly paired patches, only 22 had distances less than this threshold. In this case, the L^1 norm was used; similar results were obtained with the L^∞ and L^2 norms.

Thus, each local patch is represented by eight statistically independent features that show a strong ability to discriminate.

3.3.3 Overall Object Representation

The object is represented by a set of images taken from a sampling of views. Because of the above-mentioned feature stability, each view angle parameter (pan and tilt) is sampled at roughly 45 degree intervals, where pan is a rotation about the object's vertical axis, and tilt is a rotation away from it. For the final experiment in this section, nine view samples were used, separated by roughly 45 degrees, covering somewhat less than a hemisphere of views (approximately a 135 degree spread).

From each image, patches of different scales and positions are selected. Each patch has a scale associated with it, the *base scale*, that is the larger of the two Gaussian scales used to represent it. By using patches of different base scales, the object can be detected as it varies in size. Since the patch features used here have shown good stability for up to 20 percent scale change (approximately a quarter octave), base scales with up to half an octave of separation can be used. In the final experiment reported here, patches of three different base scales were used to cover about an octave range in scale.

With tens or hundreds of thousands of pixels in an image, the image patch positions also must be subsampled. Regular grid sampling is not necessarily desirable since the sampling in the reference and input image are bound to not match under view change.

For the experiments presented here, patches were selected based on the following criteria:

- The gradient of the larger scale must be above some threshold (contrast threshold). This allows the normalization to be stable.
- The point must be a zero-crossing of the Laplacian at the larger scale. This selects a scattering of patches at about the right density for the objects studied here (human faces), and since the zero-crossings tend to be stable as the view changes, the patch positions tend to be roughly aligned.
- A point must have patch features sufficiently different from its neighbors. (Neighbor distance threshold.)

Following these policies creates roughly 150-250 patches for a 256 by 240 image and a base scale (Gaussian σ) of 5 to 7 pixels. With three scales and nine views, roughly 4,000 to 7,000 patches are created to model the object's appearance. It is important to note that the zero-crossing restriction reduces the number of patch features to seven, since the two second derivatives at the larger scale are now dependent. However, they are effectively used to discriminate the selected patches from the great majority that are not selected.

3.3.4 Indexing

Since indexing is not the final step of detection in our design, and is only used to make the matching at the local patch level efficient, complex indexing strategies, such as in [7], have not been stressed (though they may enhance the performance). Instead, the features of a patch are quantized, and the quantized vector is used as an index. To allow for some additional feature variation, the quantization ranges overlap by some amount (quantization overlap). This means that there will be multiple entries in the table per model patch stored. However, during recognition, the input image patch indexes only a single cell and thus is very efficient.

For the final experiment presented here, a quantization level of three buckets was used. This gives us a table of $3^7 = 2187$ distinct cells. The overlap policy produces about 10 entries per stored patch, or roughly 40,000 to 60,000 entries. The retrieval rate

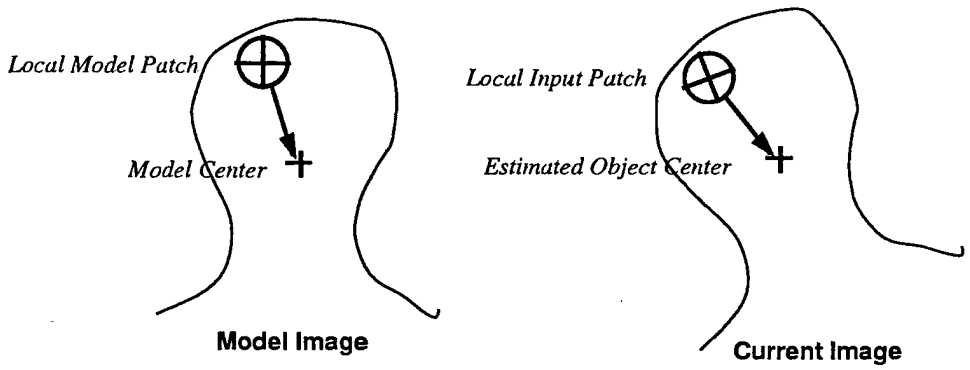

Figure 5: Pose Estimates. Given matching model and input patches, the position of the object center in the current image is estimated. It is estimated from the stored position of the center relative to the model patch and the computed transformation between the local reference frame of the model match and that of the input patch.

(the number of model patches retrieved per input image patch during matching) ranges from 30 to 70. Considering that there are 4,000 to 7,000 patches used to represent the 3-D object in a wide range of poses, this implies that, at most, one percent of the model patches are selected during indexing.

The model patches retrieved by an input patch index are then matched against the input patch. As discussed above, different norms have been compared for discrimination effectiveness under image distortion and have been found to be similar, with the L^1 and L^2 norms performing best. The L^2 norm was used for the recognition experiments discussed, though the L^1 may be more efficient on certain computers.

3.3.5 Voting and Detection

Once a model and input patch match, information about their positions, scales and orientations of their gradients, as well as information about the 3-D view associated with the model patch, can be used to roughly estimate the object pose transformation consistent with match.

Stored with the model patch is its scale and its 3-D view in pan and tilt parameters;

this allows a rough estimate of these pose parameters. The 2-D rotation between the patches can be estimated by the difference in the gradient orientation of the two matched patches. This leaves the parameters representing the object translation in the image plane.

Image translation is estimated in two steps. During model image capture, a point on the surface of the object is tracked from frame to frame and is used as an arbitrary origin for the object (Figure 5). For each model patch, the image position of the origin relative to the model patch is stored with the patch. This relative position is in a reference frame defined by the position of the model patch in the image, its base scale, and the orientation of the gradient at the base scale.

During recognition, the object origin will have approximately the same relative position with respect to the correctly matched input patch as the model patch, and the origin's absolute image position can be recovered given the position, gradient orientation and scale of the matching input patch. This object position calculation during recognition is very simple, making the overall pose voting step very efficient. The pose estimate also is fairly accurate, producing clear clusters in pose space when sufficient matches are generated. Figure 6 shows an example of this. The accumulator is shown as an image with brightness representing vote count, and the horizontal and vertical axes represent image x and y respectively. The bright spot is a concentration of votes at the correct pose.

For the final experiment presented here, the 6-D voting space included the full range of positions possible in a 256 by 240 image, 180 degrees of range for each rotation parameter (pan, tilt, and rotation in image), and an octave range of scale. The voting accumulator was quantized so that x, y, image rotation, scale and tilt had 32, 30, 4, 3 and 3 cells, respectively. Pan was given only one cell (different pans were not distinguished), to keep the overall size of the accumulator down to an efficient size.

3.3.6 Experiments with Brightness Moments

In the first experiment, a single model image is compared to a range of input images to demonstrate the stability of detection. Figure 7 shows sample images from a twelve frame sequence with changing features (eyes closing and mouth opening) and view. The

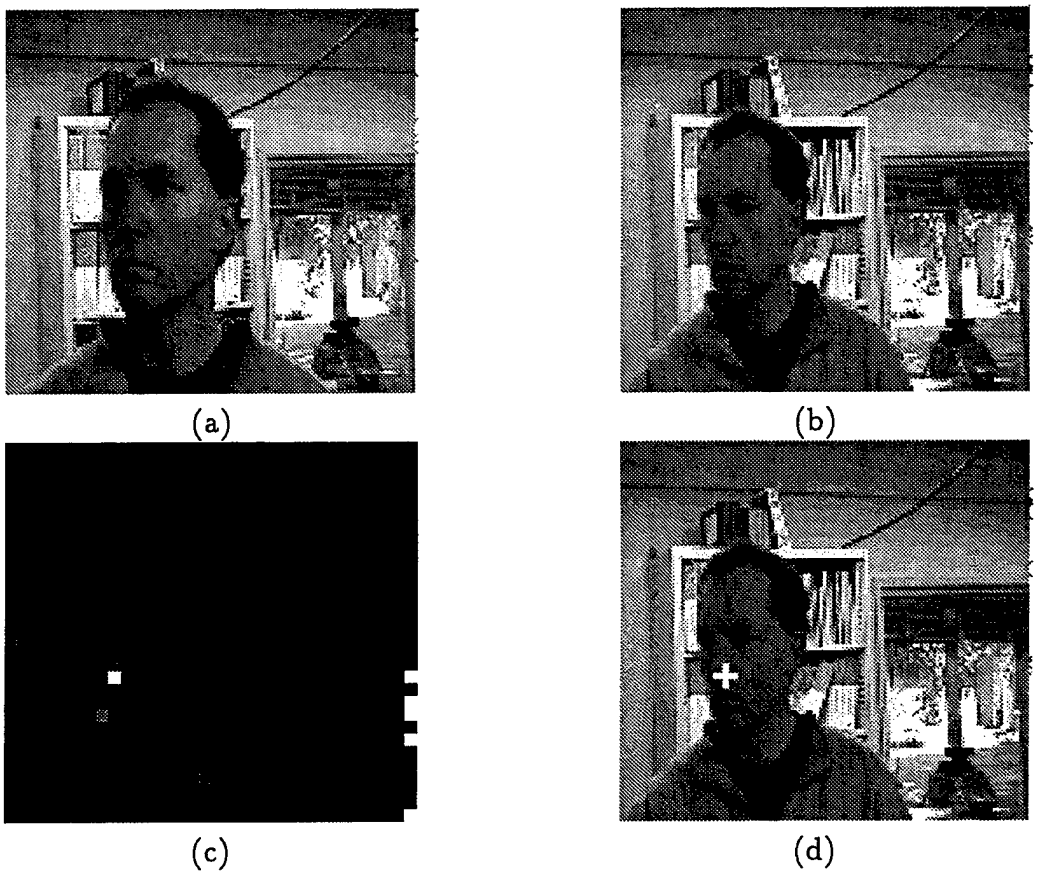

Figure 6: Detecting Clusters of Votes in Pose Space. (a) Model image. (b) Current input image. (c) Slice of pose accumulator showing vote distribution across cells representing image positions x and y (horizontal and vertical axes respectively.) Brightness of cell represents vote count, with cell of correct pose forming a clear, strong peak. (d) Resulting detected and localized object. (The subject's nose is arbitrarily used as the object origin in all experiments presented here.)

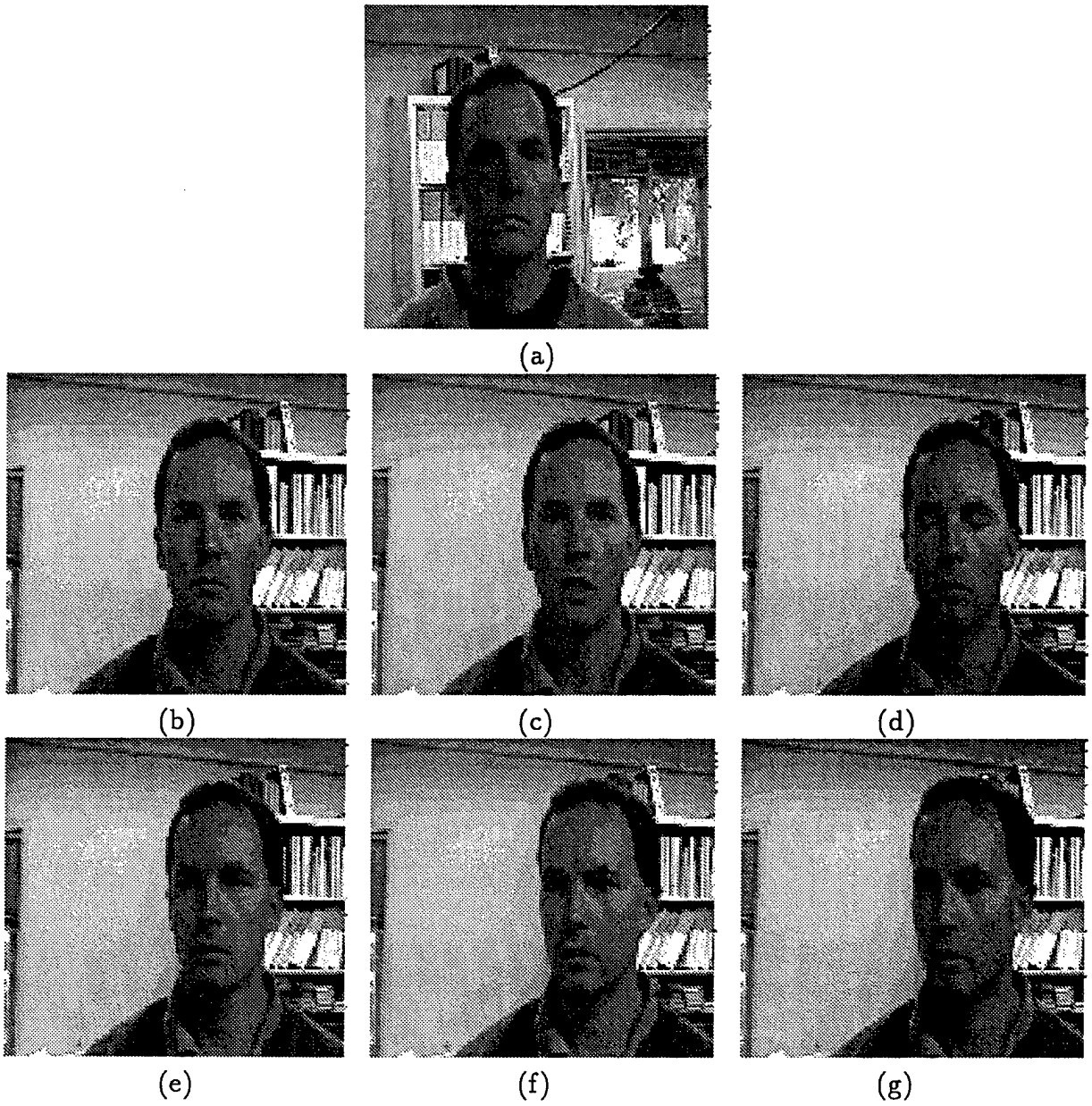

Figure 7: Test of Pose Limits with Single Model. Experiment demonstrating stability of detection using one model image. (a) Model image. (b-e) Examples of the 10 out of 12 frames that had clear, strong and correct peaks, in spite of feature movements and head rotation. (f) Frame with correct, but weaker peak. (g) Frame with over 40 degrees of rotation, some scale change and failed detection.

first ten images were strongly matched – the correct peak in the accumulator was four times higher than any random cluster, where a peak is judged correct if it is within approximately one accumulator cell of object origin's position. These ten strong matches include those to the first four images shown in the figure. In the eleventh image, the correct peak is still the highest, but is only slightly higher than the clutter. In the final image, the rotation is beyond the range of the model image and the correct peak is not selected by the matcher. However, this rotation is over 40 degrees and the scale has changed. Overall, this sequence shows the potential usefulness of the method. In these tests, the best performance was achieved when the contrast threshold was set to 5.0, the minimum neighbor distance set to 0.5, and the feature quantization overlap was set to 0.3.

In the second experiment, a person is detected in a 150 frame sequence of him talking, blinking and rotating. Three model images were used, shown in Figure 3; the four images shown in the figure display the general range of variation. The rotations included rotations in the image plane and about a vertical axis (pan) for a range of over ninety degrees. The model images were selected at three different pan positions, and the system correctly detected and localized the face in each of the 150 frames. The correct peaks were several times higher than the random clutter.

In a final brightness moment experiment, the full 6-D pose system was tested with all the parameters set as discussed above (Figure 8.). The subject was allowed large motions in all six degrees of freedom, as well as talking, blinking and other feature changes; the scale changes covered over half an octave (50% change in scale). One hundred frames were grabbed over a 20 second interval. In 98 out of 100 of the frames, the system correctly detected and localized the face, and, for an overwhelming majority, the correct peak was at least two to three times higher than any random clustering of clutter in the accumulator. For the two frames with bad matches, one had a clear, strong peak near the object origin, but not within one accumulator cell (it was more than four cells away). This may have been because of the fact that the actual object 3-D orientation was between those sampled for modeling, producing errors in the origin estimate. In the other bad match, there was a clear, correct peak, but it was not high enough above accumulated clutter in other parts of the voting space.

Figure 8: Test with Multiple Models. Experiment demonstrating the full system with multiple model images and a 100 frame sequence. For 98 out of 100 frames, the cell associated with the correct pose was the accumulator peak, and was almost always clear and strong. The nine correctly matched samples shown above demonstrate the range of pose and feature changes.

3.4 Lighting Change and Orientation Fields

For recognition to be successful, the detection rates must be stable with respect to large changes in lighting. The above brightness moment system is invariant to changes in image contrast through normalization of the features; however, it is not explicitly tolerant of large changes in the direction and distribution of light sources. Figure 9 shows the effect of such changes on the appearance of an object. Much of the brightness variation in the image of a face is in fact because of shading, and by changing the lighting, the shading, and hence, the brightness moments representing it, undergo large changes. These changes can often include complete reversals in sign.

Is it important to work with properties of the image that tend to be stable with respect to lighting changes. One property that is often stable is the general direction of the brightness variation modulo 180 degrees: even though the magnitude may vary and the sign may flip, the direction of the gradient is often constrained to the lie near a line. Figure 9 shows the orientation fields for two different lighting conditions. This stability is certainly true at the projection of many types of edges, including physical, occluding, and reflectance. In addition, this tends to be true in shading. Over a significant range of light source directions and object surface curvatures, the gradient orientation lies near a line parallel to the projection of the direction of maximum magnitude curvature of the surface being illuminated. The more extreme the ratio of principal surface curvatures, the more this is the case: for cylinders, the shading gradient is almost always so oriented. For many surfaces with more finite ratios, this is still true over a large range of light source positions.

For this reason, we developed and explored a method of using moments of the gradient orientation field (modulo 180).

3.4.1 Local Representation

The local representation requires that we compute the average gradient direction (modulo 180) within a Gaussian weighted window of variable size. One measure that has this property is the eigenvector associated with the minimum eigenvalue of the smoothed texture matrix of Lindeburg and Garding [12]:

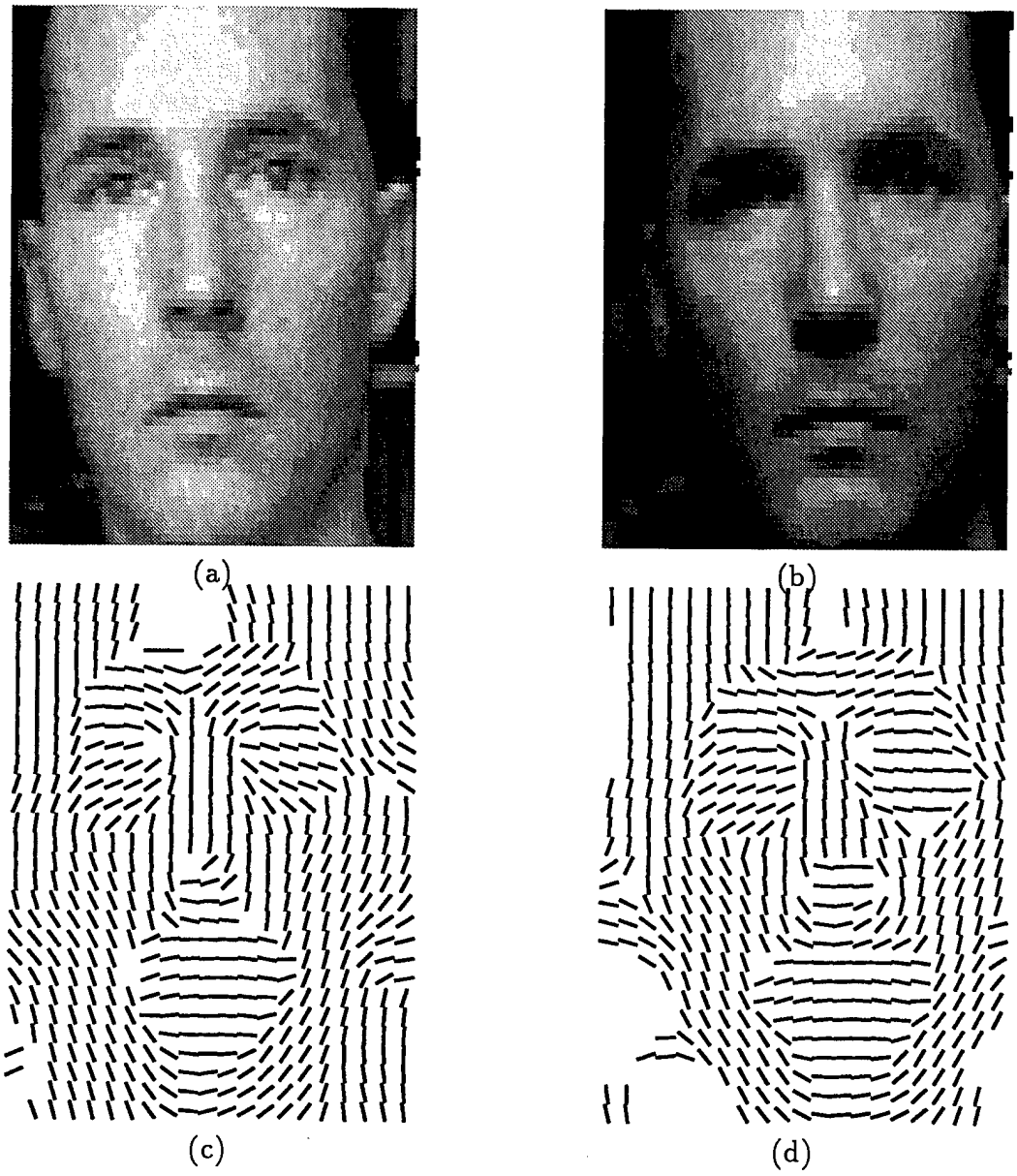

Figure 9: Local Texture Orientation Features. Example of object appearance under large lighting changes. (a - b) Two images of a face with different lighting. (c - d) The local texture orientation (normal to the gradient and modulo 180 degrees) demonstrating the potential stability of this local feature.

$$W \left| \begin{array}{cc} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{array} \right|$$

where $(g_x, g_y)^T$ is the gradient of the image convolved with the Gaussian at the *texture scale*, and W is the Gaussian weighted averaging of the matrix terms at the *integration scale*. The resulting eigenvector reflects the texture orientation θ in the underlying image, and is invariant to the sign and magnitude of the texture contrast. Figure 9 show the resulting θ fields, where the bar directions are actually aligned with the texture orientation (normal to the gradient orientation).

By choosing different integration scales and differentiating the resulting orientation fields with respect to the image x and y, we have multiscale orientation moments. Figure 10 gives a feel for the different characteristics of the fields that the different first and second derivatives are sensitive to. Each of the five fields produces a significant response from one of the derivatives, and zero from the others: (a) θ_x, (b) θ_y, (c) θ_{xx}, (d) θ_{xy} and (e) θ_{yy}, respectively.

For our experiments, we used two integration scales, an octave apart, and all first and second derivatives of the orientation at each scale. The values are normalized with respect to image rotation by rotating (steering) the direction of differentiation to be parallel x' and orthogonal y' to the orientation field at that point. Principal component analysis of the data also is done, as in the brightness data. This gives us ten normalized and orthogonal features for local patch matching.

3.4.2 Experiments in Lighting Tolerance

The detection system using orientation moments is essentially the same as the brightness system discussed above. The major difference is that the quantization level was set to two: only the sign bits of the features were used. It is unclear how stable these features are, and, since there are three more features per patch, we could afford to use less information per feature. Another difference is that the sign ambiguity of θ (modulo 180 degrees) creates an ambiguous pose estimate: two poses are consistent which each patch match, hence, two votes are generated.

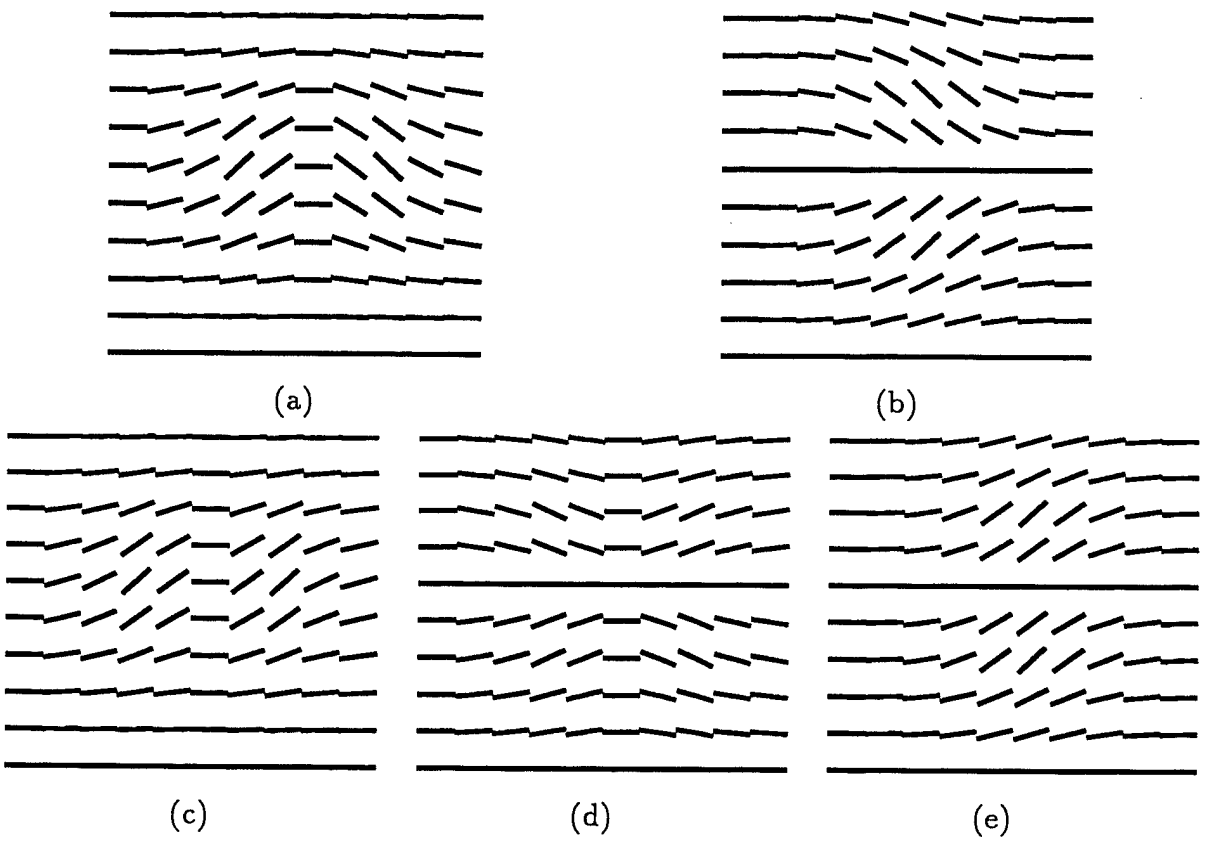

Figure 10: Multiscale Orientation Moments. Examples of orientation fields that the different first and second derivatives respond to selectively. Each of the five fields produces a significant response from one of the derivatives and zero from the others: (a) θ_x, (b) θ_y, (c) θ_{xx}, (d) θ_{xy} and (e) θ_{yy}, respectively.

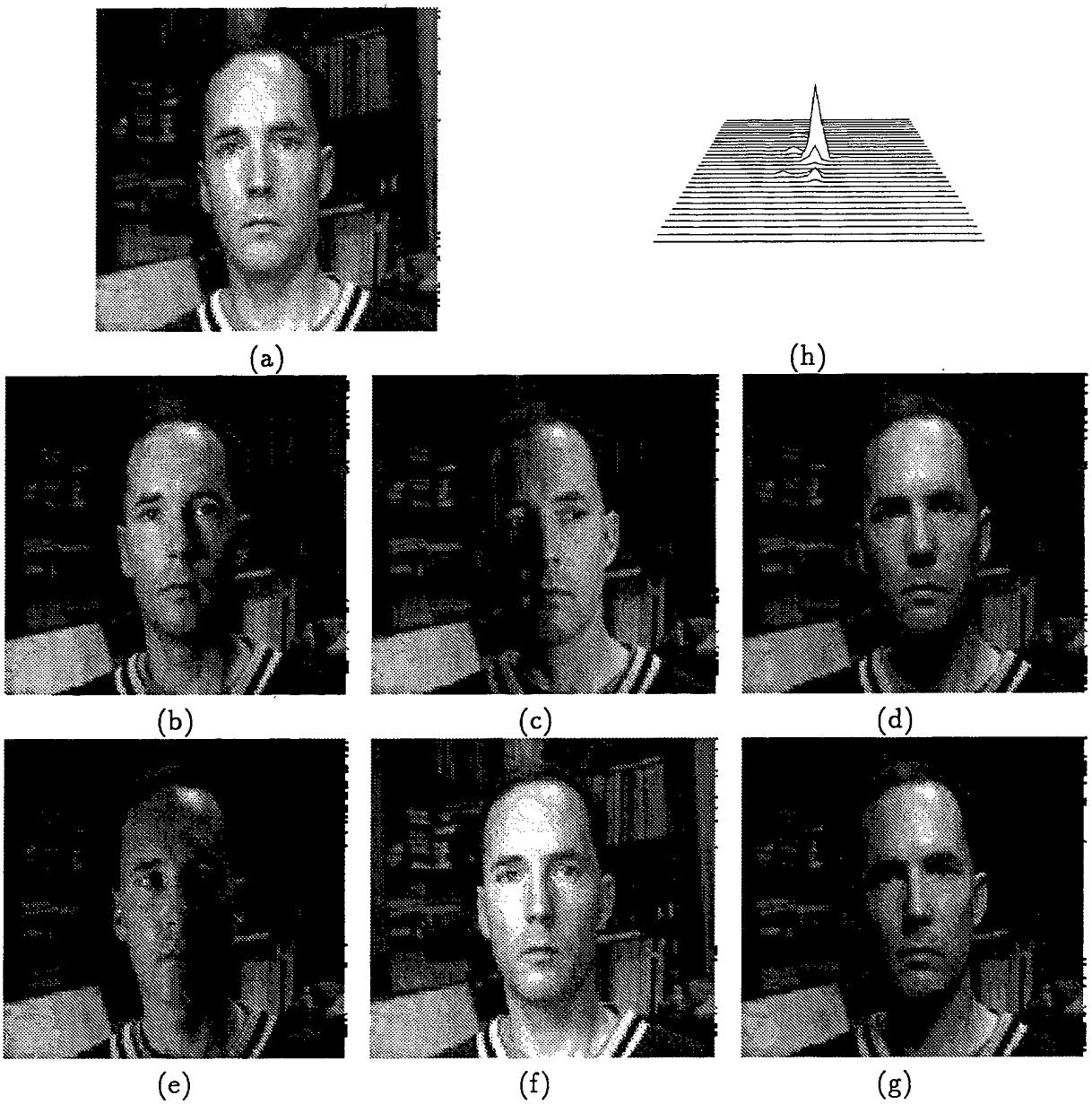

Figure 11: Recognition with Moments of Orientation. Demonstration of recognition based on moments of orientation. (a) Model image. (b-g) Six of the 25 images taken under varying lighting; 24 of the 25 were correctly matched, with clear strong peaks, including the above six. (h) The vote accumulator of image (b) showing a clear, correct peak.

Figure 11 shows the results from an experiment where the lighting was varied dramatically and everything else was held roughly constant. Twenty-five images were taken, with the light source ranging in the pan and tilt directions by more than 90 degrees. (The light was approximately a meter away.) In some frames, this concentrated light source was the sole source, while in others, a large, strong diffuse source was added (a large window). One image was selected as the model (Figure 11), and all were matched to it. In twenty-four out of twenty-five of the images (96 percent), the correct peak in the voting space was selected. It was typically strong and clear (Figure 11). In the misdetected image, the correct peak is still clearly discernable, however some spurious clutter created a higher peak elsewhere.

3.5 Conclusions on Recognition Research

This study combines two useful methods in recognition: consensus or voting-based approaches and moment-based representations. This combined method is an improvement over voting and moment methods in isolation. Using image brightness moments, the idea is successfully demonstrated on examples of human faces undergoing full 3-D pose change, as well as changes in features such as talking and blinking. The idea is then extended to moments of local texture orientation and successfully demonstrated under large variations in lighting.

Overall, the detection rates are very good for large ranges of 3-D poses. The system has the potential to be fast during recognition. Gaussian convolution can be very fast and only two scales are used. The finite differences and feature transformation required are simple and need only be performed on the roughly 1,000 patches selected per input image. In addition, for each input patch, the indexing is very fast, and with an average of 30-70 model patches retrieved, only 30,000 to 70,000 patch comparisons are performed. Future work includes a real-time implementation on a conventional computer. Achieving real-time recognition is part of the motivation for the design.

One area for continuing improvement is the height of the correct peak relative to the height of spurious peaks generated by random clutter. One method of doing this is to increase the number of features used to filter out more bad patch matches – perhaps by using the first three Gaussian derivatives. Another valuable experiment is to combine the use of both brightness and orientation moments. The former represents information

that many times can be useful (e.g., the sign of contrast), while the latter should be more robust in many other situations.

4 Report Summary

This Annual Report details progress that Teleos has made in the development of computer vision and visual attention mechanisms for the support of a S&S-directed vision and planning system.

The major visual perception capabilities relevant to security and surveillance that we addressed in this research program are the detection of human subjects, the tracking of their motion and the re-detection of specific, previously monitored subjects. Research was performed this year towards this end, and progress has been made in the following critical areas:

1. The study and development of real-time methods for discriminating moving shapes against moving backgrounds.
2. Investigation of object recognition using consensus based techniques to increase robustness and computational efficiency.

The work performed in each of these areas is summarized in the report, and an extended technical presentation is given of the research on consensus based recognition.

References

[1] D. W. Thompson and J. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *Proc. IEEE Int. Conf. on Rob. and Auto.*, pages 208–220, Raleigh, NC, 1987.

[2] Y. Lamdan and H. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *Proceedings of the Second International Conference on Computer Vision*, pages 238–249, Tampa, FL, December 5-8, 1988.

[3] A. Califano and R. Mohan. Systematic design of indexing strategies for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 709, New York, NY, June 15-17, 1993.

[4] A. Khotanzad and Y. Hong. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), May 1990.

[5] T. Reiss. *Recognizing planar objects using invariant image features.* Springer-Velag, 1993.

[6] D. Ballard and L. Wixson. Object recognition using steerable filters at multiple scales. In *IEEE Workshop on Qualitative Vision*, page 2, 1993.

[7] R. Rao and D. Ballard. Object indexing using iconic sparse distributed memory. In *Proceedings of the Fifth International Conference on Computer Vision*, page 24, Cambridge, MA, 1995.

[8] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Proceedings of the Fifth International Conference on Computer Vision*, page 637, Cambridge, MA, 1995.

[9] X. Wu and B. Bhanu. Gabor wavelets for 3d object recognition. In *Proceedings of the Fifth International Conference on Computer Vision*, page 537, Cambridge, MA, 1995.

[10] W. T. Freeman and E. H. Adelson. Steerable filters. In *Proceedings of the Optical Society Image Understanding Workshop*, volume 14, pages 114–117, North Falmouth, Cape Cod, Massachusetts, June 12-14, 1989.

[11] R. Manmatha and J. Oliensis. Extracting affine deformations from image patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 754, New York, NY, June 15-17, 1993.

[12] T. Lindeburg and J. Garding. Shape from texture from a multi-scale perspective. In *Proceedings of the Fourth International Conference on Computer Vision*, page 683, Berlin, Germany, May 11-14, 1993.